Parallel Aligned Treebank Corpora at LDC: Methodology, Annotation and Integration

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Xiaoyi Ma, Niyu Ge, Ann Bies, Nianwen Xue, Mohamed Maamouri

Linguistic Data Consortium, IBM, Brandeis University
Email:{xuansong,strassel,sgrimes,safa,xma,bies,maamouri}@ldc.upenn.edu, niyuge@us.ibm.com, xuen@brandeis.edu

Abstract

The interest in syntactically-annotated data for improving machine translation quality has spurred the growing demand for parallel aligned treebank data. To meet this demand, the Linguistic Data Consortium (LDC) has created large volume, multi-lingual and multi-level aligned treebank corpora by aligning and integrating existing treebank annotation resources. Such corpora are more useful when the alignment is further enriched with contextual and linguistic information. This paper details how we create these enriched parallel aligned corpora, addressing approaches, methodologies, theories, technologies, complications, and cross-lingual features.

1 Introduction

Parallel aligned treebank (PAT) refers to sentence-aligned data annotated with morphological/syntactic structures and aligned manually or automatically at one or more sub-sentence levels, such as the Japanese-English-Chinese PAT (Uchimoto et al. [7]) or the English-German-Swedish PAT (Volk et al., [8]). Incorporating contextual/linguistic information into a PAT is a new trend, opening up new possibilities for reducing word alignment error rate (Ittycheriah et al. [2]) and enhancing translation quality in statistical machine translation (SMT) models. One such effort is the incorporation of contextual features into tree-alignment (Tiedemann et al. [6]). As a part of this trend, LDC is now manually aligning Penn treebanks. To enrich the word-level alignment, a layer of tagging annotation is incorporated into the alignment to capture contextual and cross-lingual features. Focusing on Arabic, Chinese, and English, LDC has produced a large amount of PAT data as shown in Figure 1.

	Arabic-English PAT				**Chinese-English PAT**				
Genre	***Arb-w***	***Token***	***En-w***	***Seg***	***Ch-w***	***Char***	***En-w***	***Ctb-w***	***Seg***
NW	198558	290064	261303	8322	160477	240920	164161	145925	5322
BN	201421	259047	266601	12109	---	---	---	---	---
BC	---	---	---	---	117630	176448	91650	122714	7156
WB	19296	28138	26382	853	86263	129594	89866	82585	3920
Total	419275	577249	554286	21284	364370	546962	345677	351221	16398

Figure 1: Data Profile

In the above chart, NW, BN, BC, and WB stand for newswire, broadcast news, broadcast conversation, and web data, "Arb-w" for Arabic source words, "En-w" for English words, "Ch-w" for Chinese words, "Char" for Chinese characters, "Ctb-w" for Chinese treebank words, "Token" for tokenized tokens, and "Seg" for segmented sentences. Chinese words are based on characters/1.5.

The most common practice of creating a PAT corpus is to align existing treebank data. Such treebank resources provide mono-lingual syntactic annotations on tokens produced by a particular tokenization scheme. The alignment annotation begins with these leaf tokens to produce ground/base level alignment upon which higher-level alignment can be automatically induced. The optimal ground/base level alignment should be based on the minimum translation unit. In the context of parallel alignment, the minimum translation units refer to context-free atomic semantic units during translation. In this paper, we call it a *linear* approach if the tree leaf tokens are used as the minimum translation unit for alignment. Unfortunately, the tokens used for treebank annotation may not always be the desired minimum tokens for ground/base level alignment. Then the *non-linear* approach would call for another tokenization scheme (other than the treebank tokenization) to produce minimum translation tokens. At LDC, we create the Arabic-English PAT following the *linear* approach, and the Chinese-English PAT following the *non-linear* approach.

The paper is laid out as follows: Sections 2 and 3 discuss data source and tokenization issues respectively; Section 4 elaborates on alignment and tagging annotation at LDC; Section 5 introduces treebanks used for LDC PAT corpora; Section 6 presents the data structure of a PAT; Section 7 describes complications and challenges in creating a PAT; Section 8 concludes the paper.

2 Data Source

Source data used for PAT corpora are harvested by LDC in four genres: newswire, broadcast news, broadcast conversation, and web. Source Arabic and Chinese data are collected from various TV/broadcast programs (Figure 2). Web data are newsgroups and weblogs from on-line resources. The harvested data are manually segmented into sentences by LDC, which are further outsourced to professional translation agencies to produce high quality English translation data.

Language	***Source of Programs***
Arabic	Agence France Presse, Al-Ahram, Al Hayat, Al Quds-Al Arabi, An Nahar, Asharq Al-Awsat, Assabah, Al Alam News Channel, Al Arabiyah, Al Fayha, Al Hiwar, Al Iraqiyah, Al Ordiniyah, Bahrain TV, Dubai TV, Oman TV, PAC Ltd., Saudi TV, Syria TV, Aljazeera.
Chinese	China Military Online, Chinanews.com, Guangming Daily, People's Daily Online, Xinhua News, China Central TV, 2005 Phoenix TV, Sinorama magazines.

Figure 2: Data Sources

3 Tokenization and Segmentation

Raw data need to be tokenized and/or segmented for alignment and treebank annotation. When a PAT corpus is created with the *non-linear* approach, another tokenization scheme needs to be defined for the base-level alignment. With the *linear* approach, no further tokenization scheme is needed. Both of the approaches directly extract leaf tokens from existing parallel treebank data. The extracted tokens may or may not be the smallest translation units for alignment. For our PAT, we use the extracted English and Arabic tokens as the minimum translation units for base-level alignment while the extracted Chinese tokens cannot serve as base-level alignment tokens because some of them need to be further split in order to become minimum translation units.

The English tokens are leaves from the Penn English Treebank. The tokenization has the following features: words separated by white spaces, contractions split, punctuations separated from surrounding words, and the apostrophe (‘,‘s) treated as a separate token. Most hyphens are separate tokens while some are treated as part of words.

Arabic tokenization/segmentation is complex due to the rich morphological features of Arabic. Arabic treebank tokenization splits clitics (except “determiner”) into separate tokens, allowing for finer alignment and treebank annotation. Treebank annotation markup, such as “empty category” markers, is treated as separate tokens in the alignment annotation. Punctuation is also separated from preceding tokens.

With Chinese, segmentation is challenging due to the lack of word boundaries (Wu. [9]). Segmenting raw data into individual characters is the simplest kind of word segmentation, with each character being a token. More sophisticated segmentation schemes in MT systems group characters into words which consist of one or more characters. The word segmentation scheme proposed by the Penn Chinese treebank (CTB) team (Xue et al. [10]) is one of such schemes. We directly extract leaf tokens from the Penn CTB where the Penn CTB word segmentation scheme is applied. The extracted words are used for an intermediate alignment between character-level and larger syntactic unit alignments. To enforce data consistency and integrity, instead of segmenting raw files, we further segment the CTB-word segmentation files into character-based files, and thus following the *non-linear* approach. Each character and hyphen is a separate token, and punctuation is also separated from the preceding characters. The base-level alignment for our Chinese-English PAT begins at this character-level.

4. Alignment and Tagging Annotation

4.1 Levels of Alignment and Tagging

To build a PAT corpus, the data need to be aligned either at a specific level or at several levels. The base-level alignment is built on minimum translation units.

Upward, higher-level alignments are performed on larger linguistic units, such as tree-to-tree alignment. Generally, the base-level alignment is the *word* alignment. Arabic-English base-level alignment is at the word level. With Chinese, however, the minimum linguistic unit is a character. We chose the CTB for building the PAT, and the larger component alignment is the result of applying the CTB word segmentation scheme. Therefore, the alignment annotation at the LDC focuses on the Arabic-English word alignment, the Chinese character-level alignment, and the CTB word alignment. The first two are manual alignments while the CTB word alignment is automatically induced. To enrich the Chinese-English alignment, a layer of tagging annotation is performed manually on top of the character-level alignment and is automatically propagated to the CTB-word alignment.

4.2 Word Alignment Annotation

The task of word alignment is to identify correspondences between words, phrases or groups of words in a set of parallel texts. With reference to the Annotation Style Guide for the Blinker Project (Melamed, [5]), we developed two sets of alignment guidelines: Chinese-English and Arabic-English, which can be accessed from: http://projects.ldc.upenn.edu/gale/task_specifications/.

The guidelines discuss universal alignment approaches in addition to idiosyncrasies specific to the given language pair. General strategies and principles specify rules for annotating universal linguistic features, and specific rules are for idiosyncratic language features. The Arabic guidelines address Arabic-specific features, such as equational sentences, empty subjects, cliticization of determiners, prepositions, pronouns, and conjunctions, idioms and certain Arabic interrogative words with no equivalent words in English. For Chinese-English alignment, specific topics include the Chinese particles, non-inflection, topicalization, measure words, duplication, tense and aspects, various types of helping words.

Two types of links (*translated-correct* and *translated-incorrect*) and two types of markups (*not-translated correct* and *not-translated incorrect*) are designed to capture general linguistic information and language specific features. Most of the alignment links are *translated-correct* links which indicate valid translation pairs. *Translated incorrect* link type covers instances of erroneous translations lexically, grammatically or both. *Not-translated incorrect* refers to cases with a loss of semantic meaning and an absence of surface structure representation. For unaligned words, such as omissions or insertions of words, we use the *not-translated correct* markup to indicate cross-lingual features.

Two approaches are proposed for word alignment: *minimum match* and *attachment*. The *minimum match* approach, illustrated in Figure 3, aims to identify complete and minimal semantic translation units, i.e., atomic translation pairs. This method helps to map minimum syntactic structure unit equivalence, generating minimal semantic unit alignments which may be one-to-one, many-to-one or many-to-many links. The *attachment* approach is introduced to handle unaligned words.

The unaligned words are normally contextually or functionally required for semantic equivalence but they do not have surface structure translation equivalence. With the *attachment* method, shown in Figure 4, the unaligned words are attached to their constituent head words to indicate phrasal constituent dependency or collocation dependency. Unaligned words at the sentence or discourse level are not attached because they have no immediate constituents to depend on and attach to.

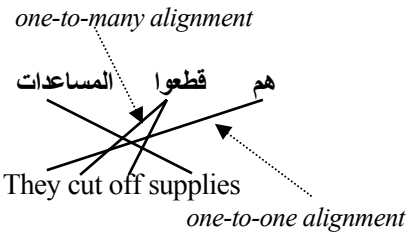

Figure 3: Minimum Match Approach

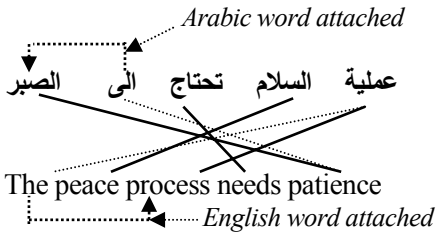

Figure 4: Attachment Approach

4.3 Tagging Annotation

To improve automatic word alignment and ultimately MT quality, researchers are exploring the possibility of incorporating extra information into word alignment. Following this direction, LDC collaborated with IBM in creating an additional layer of annotation by adding linguistic tags to the existing word alignments. Tags are added to both source and target languages to indicate different alignment types or functions of unaligned words. The tagging guidelines were jointly developed by LDC and IBM. The tags can be language independent, but the current tagging focus at LDC is the Chinese-English alignment. The Arabic alignment guidelines were updated to include a new word tag "GLU" for unaligned words, whereas for Chinese-English alignment, a set of tags were designed in the tagging guidelines for labeling all the aligned links and unaligned words (Li et al. [3]).

For Chinese-English alignment, we designed seven link types and fourteen word tags (Figures 5 and 6) to systematically address a variety of linguistic features.

Alignment Link Tags	***Examples***
Semantic	这个(this)***教授***(professor) [this ***professor***]
Function	***在***(in)这个(this)工厂(factory) [***in*** this factory]
Grammatically-inferred	***把工作***(work)完成(finish) [finish this ***work***]
Contextually-inferred	欢迎***收看CCTV***[Welcome to ***CCTV***]
DE-clause	离开(left)***的***女士(lady) [lady ***who*** has left]
DE-modifier	问题(issue)***的***(of)实质(nature) [the nature ***of*** this issue]
DE-possessive	教授(professors)***的***(from)关注(attention) [attention ***from*** the professors]

Figure 5: Link Types

Word Tags	*Examples*
Omni-function-preposition	***把***工作(work)完成(finish) [finish the work]
Tense/passive	暴露(exposed)***的***问题(issue) [the issue exposed]
Measure word	两(two)***家***杂志(magazines) [two magazines]
Clause marker	他(he)犯(made)错(mistake) [the mistake ***which*** he made]
Determiner	记者(reporter)说(said)… [***The*** reporter said…
TO-infinitive	继续(continue)工作(work) [continue ***to*** work]
Co-reference	主席(chairman)说(said)将要(would)…[The chairman said ***he*** would….]
Possessive	工厂(factory)工人(workers) [the workers ***of*** this factory]
DE-modifier	干(did)***得***快(fast) [did fast..]
Local context	欢迎(welcome)***收看***(CCTV) [Welcome to CCTV]
Rhetorical	台湾(Taiwan)学生(students)和(and)大陆(mainland) ***学生***(students) [students from mainland and Taiwan]
Sentence marker	老师(Teachers)很(very)忙(busy)***的***[Teachers are very busy.]
Context-obligatory	下雨(rains)了[***It*** rains]
Non-context-obligatory	他(He)***都***已经(already)离开(left)了[He already left]

Figure 6: Word Tags

The original alignment type *translated correct* is further classified into seven link types. The fourteen word tags are used for unaligned words. In the tagging guidelines, the Chinese 的 (DE) is a particular focus because of its complexities for machine translation (Li et al. [3]). To indicate the use of the particle 的 (DE), we tag all instances of this particle in Chinese texts by labeling them with DE-related alignment type and word tag, as illustrated with examples from Figures 5 and 6 above.

4.4 CTB Word Alignment and Tagging

The CTB word alignment is obtained from automatically transferring the manually-annotated character-level alignment. The transference merges the alignments if the CTB word has more than one Chinese character. We preserve the word tags for each individual character in this automatic alignment process. Similarly, link types are preserved to indicate the contextual information and different internal sub-part structures of CTB word alignment. Figures 7 and 8 illustrate how tags are preserved after automatic CTB word alignment. Figure 7 shows two aligned links at the character-level alignment. The Chinese token 1 (鲜) is aligned to the English token 2 (fresh), and the token 2 (花) is aligned to the tokens 1 and 3 (the flowers) (see alignment file format in Section 6). The link types are "semantic (SEM)" and "grammatically-inferred semantic (GIS)" respectively. The word tag DET is for "determiner". After the CTB word alignment processing (Figure 8), the CTB token 1 (鲜花) is aligned to the English tokens 1, 2, and 3 (the fresh flowers), and we keep both link types SEM and GIS to indicate contextual information.

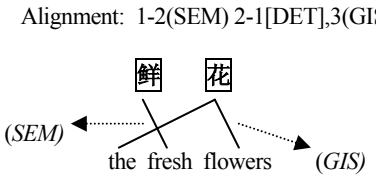

Figure 7: Character Alignment

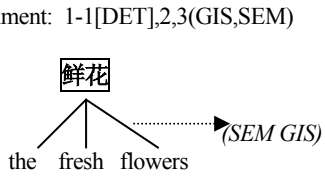

Figure 8: CTB-word Alignment

4.5 Efficiency and Consistency of Alignment and Tagging Annotation

To facilitate the annotation task, an annotation tool was developed at the LDC which allows alignment and tagging on the same interface. The annotation efficiency is monitored via the annotation workflow interface (Figure 9), where one can query the annotation volume and speed for a particular project, task, dataset, or annotator. The average annotation speed is about 8 hours per 10,000 source words for alignment and 6 hours per 10,000 source words for tagging.

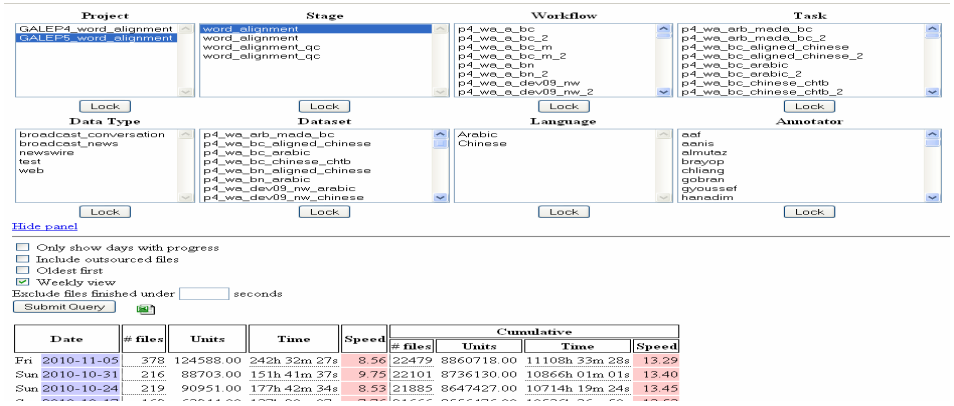

Figure 9: Efficiency Report Interface

To ensure annotation consistency, we conducted consistency tests on the pilot alignment of newswire data jointly annotated by LDC and IBM (Figure 10).

Data (Newswire)	*Chinese Characters*	*Precision*	*Recall*	*F-score*
File1	306	97.27%	95.70%	96.48%
File2	185	95.28%	96.19%	95.73%
File3	365	90.37%	91.20%	90.78%
File4	431	90.83%	92.61%	91.17%

Figure 10: Inter-annotator Agreement on Alignment

5 Treebank Annotation

Building PATs requires parallel treebanks. We use the Penn parallel treebanks for creating PATs at LDC. The Penn Arabic Treebank (ATB) annotation consists of two

phases: morphological/part-of-speech (POS) and syntactic/tree annotation. POS annotation includes morphological, morphosyntactic and gloss information. Syntactic annotation focuses on the constituent structures of word sequences, providing function categories for each non-terminal node, and identifying null elements, co-reference, traces, etc. (Maamouri et al. [4]). To build our Arabic-English PAT corpora, we started with treebank data from the most recent releases and ATB Part 3 (Bies et al. [1]). Treebank annotation markups are preserved during alignment process to maintain data integrity.

The Penn CTB corpora are segmented, POS tagged, and syntactically-annotated data. For our Chinese-English PAT corpora, we took all available CTB sources parallel to the English treebank for alignment annotation and corpora integration, excluding data with loose translations and files with improper format. The English translation treebank in correspondence to Arabic and Chinese is produced jointly by the Penn English Treebank team and the English treebank team at the LDC on four genres (BN, BC, NW and WB). For our Chinese-English and Arabic-English PAT corpora, we use English raw and tree files from the LDC published resources.

6 Data Structure and File Format

Instead of using .xml to construct the data, our PAT includes four text file types: raw, tokenized, word aligned, and treebanked data, one sentence per line without markups. Files with an identical filename base have the same number of lines, and the annotations of a specific line share the same line number. Data constructed this way is simple and straight-forward, keeping the integrity of annotation from each source while facilitating an easier annotation consistency check.

```
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (VP (ADVP (AD 4)) (ADVP (AD
5)) (VP (VA 6)))) (PU 7)))
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (ADVP (AD 4)) (ADVP (AD 5))
(VP (VV 6))) (PU 7)))
(TOP (IP (CODE 1) (NP-SBJ (PN 2)) (VP (ADVP (AD 3)) (ADVP (AD 4)) (VP (VV 5) (PU
6) (CP-OBJ-Q (IP (NP-SBJ (PN 7)) (VP (ADVP (AD 8)) (VP (VV 9) (VP (VV 10) (NP-
OBJ (NN 11) (NN 12)))))) (SP 13)))) (PU 14)))
```

Figure 11: Sample of Tree File

```
3-3(SEM) 4,5[MEA]-4[OMN],5(GIF) 6-6(TIN) -1[MET](MTA) -2[MET](MTA) 1[MET]-(MTA)
2[MET]-(MTA)
3-2(SEM) 6[CON]-(NTR) 7,8,9[DEM],10[LOC]-4[DET],5(COI) 2-3(FUN) 11,12-6(TIN) -
1[MET](MTA) 1[MET]-(MTA) 4[MET]-(MTA) 5[MET]-(MTA)
2,3-2(SEM) 7-4(TIN) 5,6-3(SEM) 4[COO]-(NTR) -1[MET](MTA) 1[MET]-(MTA)
11,12-8,9(SEM) 6[CON]-(NTR) 13,14-11(TIN) 4,5-3(FUN) 7-4,5,7(SEM) 2,3-2(FUN)
8[OMN],9,10-10(GIS) -1[MET](MTA) -6[MET](MTA) 1[MET]-(MTA)
```

Figure 12: Sample of Alignment File

The treebank and alignment files (Figures 11 and 12) do not contain token strings - only the token IDs which must be looked up in the tokenized file. Trees are represented in the Penn treebank format (labeled brackets). Tree leaves contain POS tags and token IDs corresponding to the numbers in the tokenized file. Most lines have one tree while some may have more. Multiple trees on one line are separated by whitespace. In a word alignment file, each line contains a set of alignments for a

given sentence, as shown in Figure 12, where the alignments are space-delimited, with each alignment in the format of "s-t(linktype)", *s* and *t* being a list of comma delimited source and translation token IDs respectively. The alignment type is in the parentheses and the word tags in square brackets.

7 Complications of Data Processing and Annotation

Integrating existing treebank annotation resources expedites the process of creating a PAT. However, as the down-stream annotation, the alignment process is challenging because of complications inherited from existing annotation resources.

The most common problem in data processing is segment mismatch. Mismatch may exist between source and translation raw files, between tree and raw files, and especially between translation tree and source language tree files. This problem arises when a single source sentence is translated into multiple independent English sentences. Treebank annotations of source and target all operate on single sentences. As a result, the number of source trees does not match that of target trees. We automatically re-align the mismatched sentences with an error rate below 5%. Errors resulting from this re-alignment are further handled during manual alignment annotation by rejecting the mismatched sentences. Other data processing complications include inconsistent filenames and file formats because the existing annotation resources involve different parties and various annotation stages. We standardized the filenames and converted the files into the desired release format.

Data from different sources create more noisy data for alignment annotation. Noisy data, the elements interfering normal annotation, refer here in the context of word alignment annotation to the sentences with incorrect translations/segmentations, sentences containing foreign language, or sentences that are ill-formatted. A "rejection" function is designed as a part of the alignment tool for annotators to reject such noisy data during annotation. Another type of noisy data is annotation markups carried over from up-stream annotation, for which a special tag is introduced.

8 Conclusion and Future Work

As an on-going project of the GALE (Global Autonomous Language Exploitation) program, this work has created large PAT corpora by aligning the existing parallel treebanks. Tagging annotation added to alignments is not the same as monolingual POS annotation, but rather helps to identify contextual and cross-lingual features which emerge in alignment process, thus contributing to alignment error reduction and high translation accuracy. Future efforts may scale up to richer tagging annotation, alignments of higher levels, and more language pairs.

Acknowledgement

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper

does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] Bies A., Mott J., Warner C. (2010). English Translation Treebank -- EATB Part 6 v2.0 (Annahar/ATB3 parallel). LDC Catalog Number: LDC2010E21.

[2] Ittycheriah, A. and Roukos, S. (2005). A Maximum Entropy World Aligner for Arabic-English Machine Translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 89-96.

[3] Li, X., Ge, N., Grimes, S., Strassel, S. M. and Maeda, K. (2010). Enriching word alignment with linguistic tags. In *Proceedings of LREC 2010*.

[4] Maamouri, M., Bies, A. and Kulick, S. (2008). Enhanced annotation and parsing of the Arabic Treebank. In *Proceedings of INFOS*.

[5] Melamed, D. (1998). *Annotation Style Guide for the Blinker Project* (URL: http://www.cs.nyu.edu/~melamed/ftp/papers/styleguide.ps.gz).

[6] Tiedemann, J., and Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Proceedings of TLT'08*, pp.197–208, EDUCatt: Milano/Italy.

[7] Uchimoto, K., Zhang, Y., Sudo, K., Murata, M., Sekine, S. and Hitoshi, I. (2004). Multilingual aligned parallel treebank corpus reflecting contextual information and its applications. *In Proceedings of the Workshop on Multilingual Linguistic Resources*, pp. 63-70, Geneva: Switzerland.

[8] Volk, M., Gustafson-Capková, S., Lundborg, J., Marek, T., Samuelsson, Y. and Tidström, F. (2006). XML-based phrase alignment in parallel treebanks. *In Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, pp. 93–96.

[9] Wu, D. (2009). Toward Machine Translation with Statistics and Syntax and Semantics. IEEE Automatic Speech Recognition and Understanding Workshop*,* Merano: Italy.

[10] Xue, N., Xia, F., Chiou, F. and Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207-238.